Why Hunger Is Not A Desire

Patrick Butlin

Abstract

This paper presents an account of the nature of desire, informed by psychology and neuroscience, which entails that hunger is not a desire. The account is contrasted with Schroeder's well-known empirically-informed theory of desire. It is argued that one significant virtue of the present account, in comparison with Schroeder's theory, is that it draws a sharp distinction between desires and basic drives, such as the drive for food. One reason to draw this distinction is that experiments on incentive learning show that desires and basic drives influence action in different ways.

Keywords: desire, drive, hunger, incentive learning, goal-directed behaviour

1. Introduction

> It cannot be said, however, that hunger *is* the desire or appetite for food. After consuming three-quarters of a delicious meal, I may still have an appetite – a desire to eat – even though I am no longer hungry. Furthermore, hunger typically *explains* a desire for food. But there are other possible explanations… and explanation is irreflexive. Finally, a being without the concept of food (such as a newborn baby) could perfectly well be hungry yet not be seriously said to have a desire specifically for food.
>
> - from Wayne A. Davis (1986), 'Two senses of desire', p. 65.

In this passage, Wayne Davis argues for the claim that hunger is not a desire, and his arguments get close to the heart of the matter. In this paper I will defend Davis' claim, by setting out an account of the nature of desire based on evidence from psychology and neuroscience, according to which hunger, thirst, and other *basic drives* (as I call them) are not desires. This is a particularly noteworthy feature of my theory, since it is one major way in which it differs from Timothy Schroeder's 'Reward Theory of Desire' (2004, p. 131), which is by far the best-known and most detailed empirically-informed account of the nature of desire in the philosophical literature.

Existing theories of desire often give necessary and sufficient conditions for a state's being a desire, which describe relationships with other psychological states or processes, or at least hint at such conditions. For example, Smith (1987, 1994) argues that desires are mental states with the world-to-mind direction of fit, which he takes to amount to a functional role in which the disposition to act so as to bring about the object of desire is central. Desires are therefore defined by their role in motivation. Similary, Schroeder specifies a role in reinforcement learning which he claims is constitutive of desire. But my theory takes a different approach. I assume that if there is at least one natural kind of

psychological state which fits the way that philosophers and the folk think of desires sufficiently well, then what it is to be a desire is to be a member of whichever natural kind best fits the existing philosophical and folk conception. So desires are not characterised on my view by their role in motivation, evaluation, reinforcement learning or the generation of pleasure, but by their membership of a particular natural kind. Psychological natural kinds, as I think of them, are sets of entities with common teleofunctional roles in psychological mechanisms.

My primary aim in this paper is therefore to identify and describe the most plausible candidate natural kind for this role. Natural kinds are often sub-kinds of broader kinds, and an important task for achieving this aim will be defending the relatively narrow conception of desire that I favour against broader alternatives. In particular, I will argue that the natural kind that is constitutive of desire excludes basic drives such as the drive for food, but it might be contended that Schroeder's theory identifies a broader natural kind that includes basic drives.

This may make it sound as though my dispute with Schroeder is purely verbal – that we disagree only about which of two sets of psychological states should be labelled 'desire'. But there are several respects in which my project in this paper goes beyond engagement in a verbal dispute. Most importantly, there are significant differences between basic drives and the states that I call 'desires', and we need to recognise these differences, whatever terms we use to do so. Also, as I will explain in section 5, Schroeder's account of how desires cause action is questionable (so I do dispute some of his empirical claims), and my theory emphasises points to which Schroeder gives little weight, such as the distinction between standing and occurrent desires. More broadly, I aim to contribute to philosophers' understanding of the science of motivation and action, and hence to the development of philosophical theories that are strengthened by their coherence with psychology and neuroscience.

In section 2 I give an initial presentation of my theory of desire, and in section 3 I discuss how desires are updated. My central argument for distinguishing basic drives from desires comes in section 4. In section 5 I compare my theory with Schroeder's and argue against his view, and in section 6 I offer a further argument for excluding basic drives. I conclude in section 7, pulling together the threads of the argument to explain why hunger is not a desire.

2. A Natural Kind Theory of Desire

In this section, I will present and provide initial arguments for my natural kind theory of desire. I start by identifying the subject of my theory.

My theory concerns intrinsic desires, as opposed to other pro-attitudes (Schueler 1995, Schroeder 2004). There are two important distinctions here. First, intrinsic desires are contrasted with instrumental desires, which are psychological states specifying goals that we adopt only because we believe that they will contribute in some way to our achieving further ends. Instrumental desires are produced by rational deliberative processes, whereas intrinsic desires are inputs to such processes. Second, 'desires proper' are a subset of the pro-attitudes, which include all psychological states that are capable of motivating us to act, either alone or when combined with suitable beliefs. These may

include instrumental desires, emotional urges, habits, intentions, resolutions, and normative and evaluative beliefs, and perhaps other states, as well as the intrinsic desires which are my topic.

When philosophers distinguish between intrinsic desires and other pro-attitudes, they reveal an idea which is central to the current philosophical conception of desire. This is that we are often motivated to act by the desire for some outcome, together with the belief that by acting in this way we will make the outcome more likely, but that this is only one of a number of ways in which we may be motivated to act. Acting from desire is often contrasted with acting out of habit, where habits are (roughly) overlearnt behaviours which are performed automatically in certain circumstances; and also from acting in expression of emotion (Hursthouse 1991), or acting on a resolution, formed as a result of reasoning about what is best or what one ought to do (Bratman 1987, Holton 2009). So when looking for a natural kind that fits the philosophical conception of desire, we are looking for states that motivate us in combination with instrumental beliefs; that are distinct from habits, emotions and resolutions; and that are inputs to, and not outputs from, rational deliberative processes. Philosophers also typically take us to have desires for a wide range of objects – for instance, one might desire a refreshing glass of juice or to become an astronaut – and think of desires as having occurrent and standing forms.

In behavioural psychology and cognitive neuroscience, a successful research programme distinguishes between two systems for action selection, which are called the *habit system* and the *goal-directed system* (Balleine & O'Doherty 2010). These systems are the products of two different forms of instrumental conditioning. The habit system is a mechanism by which rewards cause animals to learn associations between stimuli and responses, which in turn cause the performance of those responses when the stimuli are encountered in the future. The goal-directed system, in contrast, keeps track of two relationships: the causal relationships between actions and outcomes, and the reward values associated with outcomes.[1] It combines states representing these relationships to calculate the expected reward values of salient actions, and causes the actions with the highest expected values. The two systems are thought to exist both in humans and in many other animals. On the face of it, the states tracking the reward values of outcomes in the goal-directed system are therefore a good candidate for the natural kind we are seeking (c.f. Heyes & Dickinson 1990). These states are often referred to in the scientific literature as 'outcome values'. And indeed, this is the view that I will defend:

> *Natural Kind Theory of Desire:* For a state to be a desire is for it to be a member of the natural kind with the function of acting as inputs to the goal-directed system, which track the reward values of outcomes.[2]

[1] In this paper I take *reward value* to be a property of outcomes (i.e. states of affairs), which comes in degrees, and which in some sense makes those outcomes worth working for. Actions also have reward values derivatively, in virtue of their tendencies to produce outcomes. I also use the term 'reward' as a noun for rewarding outcomes. There is much more to be said about the nature of reward value, and how this property relates to the two systems for action selection, but here I leave these issues aside.

[2] It may be helpful to clarify my use of the term 'state', since in the empirical disciplines to which this article relates this term is typically used to refer either to states of the environment or physiological states of the organism such as the state of food-deprivation. I use 'state' in the standard philosophical way, which is to refer to individual representations in psychological systems, at either the personal or subpersonal level.

An important feature of this theory is that psychological states may sometimes represent the reward values of outcomes without being members of this natural kind, and hence without being desires. For instance, conscious explicit beliefs about the reward values of outcomes are not desires.

I now turn to evidence that the goal-directed and habit systems are distinct; that they exist in both rats and humans; and that outcome values fit the philosophical conception of desire which I just outlined.

The primary source of evidence for goal-directed control in rats is studies of *outcome devaluation* (Adams and Dickinson 1981). Outcome devaluation experiments typically have the following form. In the first phase, rats are given the opportunity to press a lever, and are given a food reward such as sucrose when they do so. The rats are exposed to this environment for long enough to learn an association between lever-pressing and reward. Then in the second phase, the rats are divided into two groups. One group experiences outcome devaluation, meaning that they are allowed to consume the particular food being used in the study before and after being injected with lithium chloride, which induces gastric illness. The second, control group also receives both the food and the injection, but these are given at different times, with the intention that the rats will treat them as unrelated events. All the rats are kept away from the lever. Finally, in the third phase, the rats are again given the opportunity to press the lever, but this test is conducted *in extinction*, meaning that the rats do not receive a reward for lever-presses. Adams and Dickinson and subsequent investigators have found that the rats for whom the outcome was devalued press the lever significantly less than the controls in this phase.

In these studies, the two groups of rats perform differently in the third phase, and this difference must be explained by some difference in the rats' experiences during the experiment. The only such difference comes in the second phase, when one group experiences the pairing of the food reward with illness, and the other does not. It is hard to resist the conclusion that the rats behave differently after this experience because they represent the food as having a different reward value, and if this is right, it must also be the case that the rats anticipate receiving this food when they press the lever. In other words, the rats' behaviour is controlled by states apparently representing the reward values of outcomes, and further states that seem to represent relationships between actions and outcomes.

Two further points about outcome devaluation studies are particularly noteworthy. First, overtraining leads to loss of sensitivity to devaluation (Adams 1981), suggesting that the rats have acquired the habit of pressing the lever. It is therefore thought that rats have distinct habitual and goal-directed systems, working in parallel, with the habit system learning more slowly. Second, humans behave in similar ways to rats in studies of this kind, including showing insensitivity after overtraining. Tricomi et al. (2009) performed a similar outcome devaluation study on humans to ones which had earlier been performed on rats (Balleine & O'Doherty 2010), with similar results. A recent experiment gives a vivid illustration of the two systems at work in humans (Neal et al. 2011). Participants were in a cinema, watching a film, and were given popcorn that was either fresh and delicious or stale and unpleasant. Those who had frequently eaten

popcorn in cinemas before ate the same amount of popcorn, even if it was stale, whereas those were not regular popcorn-eaters ate far more of the fresh popcorn. That is, those who were not 'in the habit' of eating popcorn were sensitive to whether eating the popcorn produced a pleasant outcome, but those who were 'in the habit' ate on regardless.

The evidence I have presented so far suggests that the goal-directed and habitual control systems are distinct; that they are present in humans as well as in rats; and (to a more limited extent) that desires are states of the goal-directed system. Evidence from neuroscience further supports all three of these claims. On the distinctness of the two systems, the most telling results are from a series of studies by Yin and colleagues. This group found that lesions of the posterior dorsomedial striatum (DMS) inflicted either before or after training that would otherwise produce goal-directed behaviour left rats insensitive to both outcome devaluation and *contingency degradation*, another test for goal-directed control (Yin et al. 2005). They also found that lesions of the posterior dorsolateral striatum (DLS) made rats more sensitive to these tests, when trained in ways that normally promote habitual behaviour (Yin et al. 2004, 2006). So it appears that the DLS supports habitual behaviour, that the DMS supports goal-directed behaviour, and that these two systems are independently capable of causing action. More specifically, the DMS is thought to be involved in the formation and storage of representations of relationships between actions and outcomes, while the ventral striatum has been associated with outcome values in other lesion studies (e.g. Corbit et al. 2001).

In humans and other primates, the most striking results which are relevant to this issue come from studies of the orbitofrontal cortext (OFC). These fit neatly with the studies on the striatum in rats, because the striatum and cortex are connected by partially-segregated loops that also pass through the thalamus, these corticostriatal loops are widely thought to be functional units, and orbitofrontal cortex is connected in this way with the ventral striatum (Balleine & O'Doherty 2010, Haber & Knutson 2010). Several recent reviews have discussed evidence linking the OFC with outcome values (Rangel & Hare 2010, Kennerley & Walton 2011, Padoa-Schioppa 2011). Well-known studies of brain-damaged patients by Damasio (1994) show that OFC and nearby cortical areas are necessary for normal decision-making, and single-cell recording studies in primates have found that cells in the OFC encode the identities and reward values of stimuli (Rolls & Grabenhorst 2008). Imaging studies in humans have found that OFC activity is correlated with the amounts participants are willing to pay for available goods (Plassman et al. 2007), and several studies have found OFC activity in response to a wide range of rewarding stimuli. These include attractive, smiling faces (O'Doherty et al. 2003), aesthetically pleasing paintings and music (Kirk et al. 2009), and monetary gains and erotic stimuli (Sescousse et al. 2010). This evidence suggests both that the goal-directed system exists in humans as well as rats, and that outcome values really are desires.

A further common philosophical claim about desires is that they come in two forms: *standing* and *occurrent*. Occurrent desires exist for short periods of time, are typically thought of as conscious states, and make direct contributions to determining how we act, while standing desires are persisting, non-conscious mental states that explain the recurrence of occurrent desires. This distinction is not only compatible with my

proposal, but essential to a proper understanding of desire. Two features of this distinction are particularly important in the present context.

First, it is only occurrent desires that contribute directly to action-selection. According to a popular general picture, the brain resolves uncertainty through competitions between patterns of activity associated with different possibilities (see e.g. Cisek 2007, Clark 2013, Redgrave 2007). In goal-directed action selection, it seems that instances of neural activity in the OFC and ventral striatum represent the reward values of outcomes, and contribute to the generation of patterns of activity elsewhere that represent the expected reward values (or perhaps probabilities) of possible actions. These instances of activity are plausibly occurrent desires, while the persisting structural features that govern them, for example by causing them to occur when desired outcomes are represented as possible, are standing desires. At any given time we have occurrent desires only for those outcomes that we actively represent, and the strengths of these occurrent desires may be affected by the attentional salience of those outcomes (Hare et al. 2011). So desires typically motivate us in the following way: first, we come to represent some outcome as possible or available, either because we associate it with an action that we represent as available, or because we associate it with some feature of the perceived or otherwise represented environment; this, together with our standing desires, causes the generation of an occurrent desire for the outcome; and this occurrent desire contributes to motivation to perform associated actions, in combination with occurrent representations of the probabilities of the outcome given those actions.

Second, one reason why occurrent desires are useful is that they can vary in strength dramatically, depending on the circumstances, to reflect the reward values that outcomes have under those circumstances (Padoa-Schioppa 2011, Holton & Berridge 2014). This variation may be caused in some cases by associative connections between desires and states representing levels of physiological need. For example, when an animal is dehydrated, it may be extremely physiologically important for it to consume water, and hence to have a strong desire for water. But it would be highly maladaptive for many animals to have a strong desire for water at all times, because this would prevent the pursuit of more valuable goals, and because overconsumption of water can be dangerous. Standing desires are also useful, because it is also necessary for us to keep track of the average values that outcomes have provided, on the occasions when we have experienced them – and these will be relatively stable for familiar outcomes. So it is important that we can be motivated to pursue particular outcomes to dramatically varying degrees, while maintaining stable representations of the typical long-term reward values of those outcomes, and having both standing and occurrent desires makes this possible.

Taken together, the evidence and arguments presented in this section indicate that the states with the function of tracking the reward values of outcomes in the goal-directed system form a natural kind which may be identified with desire. These states combine with others representing relationships between actions and outcomes to cause actions; they exist in humans and many other mammals; they have a wide range of objects, at least in humans; and they come in standing and occurrent forms.

3. Desire-Updating

One very important question raised by the account of desire that I have offered so far concerns how desires (i.e., outcome values) are formed and updated. It is particularly relevant here because the drive for food, among other *basic drives*, plays an important role in desire-formation and -updating. So in this section I give an initial account of desire-updating.

One possible theory of desire-updating relies on the idea that bursts and pauses in the release of dopamine, called *phasic* dopamine signals, constitute *reward prediction error signals* (RPEs). Reward prediction error signals represent the difference between the level of reward currently being received from the environment,[3] and the level that was expected to be received at this time. According to one popular account of the function of phasic dopamine signals, they take this form and are used for habit learning (Schultz 1998, Wise 2004). Schroeder incorporates this claim in his theory of desire (see section 5). Holton and Berridge (2014), meanwhile, argue that such signals update desires in the following way: desires are strengthened when the outcomes which are their objects are actively represented at the times when positive reward prediction error signals occur, and weakened by negative reward prediction error signals.

However, while this account is compatible with the theory of desire I have so far put forward, the current state of the empirical literature does not allow me to endorse it in such a precise form. There are two problems. First, there is considerable debate about the function of phasic dopamine signals; Berridge (2007) offers the most prominent dissenting view. Second, even if phasic dopamine signals do represent reward prediction errors, it appears that signals of somewhat different kinds are necessary for habit learning and for desire updating. In short, habit learning seems to be best served by signals representing the difference between the current level of reward and that expected given the action just performed (Sutton & Barto 1998), while desire learning apparently requires the difference between current reward and the level of reward previously associated with current outcomes (Holton & Berridge 2014).[4] These problems mean that any detailed account of desire-updating that I might offer would be excessively speculative.

I will therefore leave aside much of the detail of the proposal concerning reward prediction errors, and consider only a central principle of this proposal. This principle is that desires are updated by a signal that is attuned to reward in general, as opposed to by a number of distinct signals that represent specific kinds of reward, such as the satisfaction of particular physiological needs.

If this principle is correct, there must be some mechanism by which the brain evaluates the level of reward presently being provided by the environment, in order to generate reward signals (to calculate a reward prediction error, for instance, one must have available both a representation of predicted reward, and a measure of actual current reward). As Schroeder suggests, it is likely that desires themselves will play a central role

[3] i.e. the sum of the reward values of features of the environment.

[4] See Butlin 2016, for discussion and for more detail on many related issues. For brief comments on dopamine and outcome value-updating, see also Balleine et al. 2008.

in this process, since they apparently represent the reward values of outcomes. If so, this would explain the phenomenon of *secondary reinforcement*, noted by behaviourists such as Skinner (1938) and Hull (1943), in which actions are reinforced by previously meaningless stimuli such as lights or tones which have been associated with the satisfaction of basic drives.

However, to get the process of desire-updating started, we must be innately disposed to treat some outcomes as having positive reward values for the purpose of generating reward signals.[5] These outcomes are likely to include consuming palatable foods, drinking water, having sex (when mature), undergoing positive social interactions, getting warmer when cold and cooler when hot, and possibly consuming certain kinds of foods, such as sugar (Foddy & Savulescu 2010). So if the present account is correct, then our desires will typically be strengthened when we encounter their objects together with either these innately-valued outcomes, or the objects of our other existing desires. This account has several attractive features.

One attractive feature of this account is that it gets the level of responsiveness to reasons of desire-updating about right. The function of desires is to track the reward values of outcomes, so desire-updating must be responsive to some extent to evidence concerning these values. The goal-directed system would be far from adaptive if we desired at random. But we also know that the strengths of our desires do not always match our conscious, explicit judgments about what is good for us, biologically or otherwise. For example, the testimony of a doctor might immediately convince someone that eating a certain food is bad for them, but it would not immediately extinguish the desire for that food. The present account claims that desires are responsive only to evidence of the right form, and that the origins of our standing desires may be opaque to us and reach back deep into our pasts, but nonetheless that these desires will change over time. All of this seems correct.

Also, the account can readily accommodate an explanation of how we come to have the wide range of intrinsic desires that we typically attribute to one another. It is perhaps surprising that a mechanism that we share with rats and mice could explain our subtle aesthetic preferences and often abstract, specific ambitions, but the present account puts no limits on what can be desired except the agent's representational capacities. This follows from two points: first, that it is not the outcomes we *actually* experience that we come to desire, but those that we *represent* ourselves as experiencing; and second, that we are capable of off-line desire updating, by imagining outcomes, and associating them in our imagination with the objects of basic drives or existing desires. The first point I take to need no defence, and there is some empirical evidence in favour of the second. Most notably, one study has found that artificially stimulating dopamine using the drug L-DOPA while participants imagined possible future events led to them predicting greater pleasure from those events (Sharot et al. 2009). Also, it has been found that subjects are able to cause increased activity in their own midbrain dopamine neurons by imagining pleasurable scenarios (Sulzer et al. 2013), and that the OFC is activated by both real and imagined rewards (Bray et al. 2010). While this evidence is not conclusive, the claim that

[5] The use of the notion of innateness in cognitive science has been criticised (e.g. Griffiths 2002, Mameli & Bateson 2011); however, I explain and defend my use of this notion in section 6.

imagination can drive updating is clearly compatible with the reward-signal account. Such a process could help to explain how it is possible for us to have desires for outcomes we have never experienced, like becoming a surgeon or climbing a distant mountain.

With these points in mind, humans may be expected to have many, varied and sophisticated desires. We have sophisticated representational capacities; we can imagine outcomes in detail without having experienced them before, and we often exercise this ability; it is plausible that we have basic drives for social status and evidence of approval by those around us; and we live in remarkably rich and complex cultures. We are also frequently exposed to advertising designed to exploit weaknesses of our desire-formation systems, to motivate us to work for things we may or may not otherwise desire. When we add these features of human psychology to the simple general-purpose system the present account describes, it is likely to result in our having many and varied desires. This point further supports the claim that what I am calling 'desires' really are desires.

This view of desire-updating also facilitates an attractive theory of drug addiction, since many drugs of abuse boost levels of dopamine in the brain. So it may be that these drugs are addictive because they hijack the desire-updating system, causing desires for their consumption to become stronger every time they are used (Butlin & Papineau 2016).

4. Basic Drives and Goal-Directed Behaviour

Now that I have presented my theory of desire, we are close to being ready to consider the evidence that, in my view, shows most clearly that hunger is not a desire. I turn to that evidence in the second part of this section. First I discuss hunger in the context of my concept of a *basic drive*.

The drive for food in animals that are capable of reward learning plays two crucial roles. First, as I mentioned in section 3, there must be some outcomes that we are innately disposed to treat as having positive reward values. There are persisting features of our minds that cause more strongly positive reward signals to be generated, other things being equal, when we represent the occurrence of these outcomes. By analogy with standing desires, I will say that we have *standing basic drives* for these outcomes (standing desires are also persisting features that contribute to the generation of reward signals). Second, because the physiological need for food fluctuates dramatically, there must be brain states that mediate the influence of these fluctuations on behaviour, causing us to give higher priority to obtaining and consuming food when it is more urgently needed. Since food is one of several physiological needs, there are a number of drives that play this role, and hence a number of such brain states. Like occurrent desires, these brain states are instances of activity that affect action selection in the moment, so I will call them *occurrent basic drives*.

It is worth noting that there are a couple of ways in which the relationship between standing and occurrent basic drives may not mirror that between standing and occurrent desires. First, standing desires cause occurrent desires, and the same may not be true in the case of basic drives. Second, it may not be the case that every standing basic drive has an occurrent form, because not every basic drive is for something that is needed more at

some times than others – the drive for positive social interactions, for example, may have no occurrent form.

In this context, we should think of hunger as the occurrent basic drive for food. Hunger is an occurrent state that we enter when our physiological need for food is relatively high, that affects action selection in the moment. Philosophers sometimes describe hunger as a bodily sensation (e.g. Hall 2008), but we should be cautious on this point, because everyday experience suggests that it is possible to be hungry without consciously experiencing the sensation of hunger. So I leave open the question of whether hunger is identical to any particular bodily sensation. I turn now to the relationship between hunger and desire.

A first, simple reason for doubting that hunger is a desire is that there could be animals that do not possess the goal-directed system but are capable of hunger. A probable example of such an animal is the sea slug *Aplysia californica*, since these creatures are capable of habit learning and undergo changes in their motivational state depending on satiation (Brembs et al. 2002, Jing et al. 2007), but the mere conceptional possibility of creatures like this is enough to make the point. If hunger were a desire, then there could be animals that had desires, but were only capable of performing actions that were either unlearned or habitual. This point is particularly compelling if we assume, as I have, that desires form a natural kind. How can animals such as these possess states of the same natural kind as us, if they do not even possess the psychological systems within which those states operate?

A possible response to this argument is that animals that lack goal-directed systems must nevertheless behave differently depending on their occurrent basic drives. So there must be some way in which their behaviour depends on what they 'want'. However, Niv and colleagues (2006) have argued convincingly that occurrent basic drives influence habitual action by taking the role of stimuli. This means that they do not behave like desires: for instance, rats that have been trained when sated to lever-press for sugar solution do not increase responding when they are thirsty, and rats trained when hungry to perform the same action reduce responding when thirsty but not hungry, even though sugar solution is good for thirst (Niv et al. 2006). So habits are not generally sensitive to the occurrent strengths of relevant basic drives.

The role of occurrent basic drives in goal-directed control, meanwhile, is subject to a process known as *incentive learning*. As several studies have shown, goal-directed behaviours are only affected by states such as hunger when the individual has learnt about the relationship between these states and the specific outcomes they expect the behaviours to produce. Perhaps the most famous study showing this effect, by Dickinson and Dawson (1988), also provides support independent of outcome devaluation experiments for the claim that rats represent the outcomes of their actions. In this experiment, hungry but not thirsty rats were trained to perform two different actions for different food rewards. They pressed a lever to receive food pellets, and pulled a chain to receive sucrose solution. The rats were then given the opportunity to perform these actions when thirsty. Rats that had previously consumed sucrose solution when thirsty preferentially pulled the chain in extinction, but those that had not had this experience performed the two actions equally. The fact that the thirsty, experienced rats pulled the chain more than they pressed the lever shows that they represented the

outcome of this action. But for present purposes the point is that only the experienced rats showed this effect. Those that had not consumed sucrose solution when thirsty did not preferentially pull the chain, even though they associated this action with getting the sucrose solution, and had previously consumed it and found it rewarding. Simpler studies have also obtained similar results: rats that have been trained to press a lever for food will not increase their performance when hungry, compared to controls, unless they have previously consumed food of that kind when hungry (Balleine 1992, Niv et al. 2006).

What these experiments seem to show is that occurrent basic drives such as hunger control goal-directed action only indirectly. A rat that has an occurrent desire for some outcome, such as a specific kind of food, will tend to perform actions that have lead to that outcome in its experience. But a hungry rat may well fail to perform actions that have lead to its getting food in the past, if it was not hungry on those past occasions. This is because its hunger will affect its goal-directed behaviour only by strengthening its occurrent desires for specific outcomes that it has found to be rewarding, on past occasions on which it has been hungry. So hunger plays a different role in goal-directed behaviour from desires for specific foods.

A further, particularly clear demonstration of this point comes from a contrast between outcome devaluation studies that devalue the outcome using specific satiety (e.g. Balleine & Dickinson 1998) and experiments on incentive learning that involve a transition from hunger to satiety (e.g. Balleine 1992). In both paradigms, rats are given the opportunity to press a lever to receive a novel foodstuff, and learn to perform this action, then are taken away from the lever and fed until they are sated, before being given the opportunity to press the lever again without any food being delivered. The only difference is that in the outcome devaluation experiment, the same food is used that the rat has learnt to get by pressing the lever. Surprisingly, rats that undergo these procedures will press the lever less when they have been fed to satiety on the same food, and will not reduce responding when they have been fed to satiety on a different food. This seems to show that their behaviour was driven *directly* only by the desire for the specific food, and not by hunger.[6]

It is perhaps unlikely that results like those presented in this section could be obtained in adult humans. For example, it is unlikely that adult humans who were introduced to a novel food when sated, would not be more motivated to work for it when hungry. So incentive learning for specific foods may not be necessary in humans as it is in rats. But even if this difference between humans and rats is real, the best explanation for it is unlikely to be that hunger works differently – is connected to the goal-directed system in a different way – in humans as compared to rats. A more plausible explanation is that humans employ the concept of *food*, as opposed to merely representing foods of specific

[6] An apparent exception to these general claims about the role of basic drives is salt appetite. Rats that have experienced actions leading to the delivery of salt into their mouths will perform these actions when deprived of salt, even though salt delivery has previously always been an aversive experience for them (Tindell et al. 2009). There is no need for incentive learning. This suggests the existence of a special-purpose mechanism for tracking the presence of salt and responding to salt appetites, which would make some sense given that in contrast to other foods, salt appetite is a rare condition and overconsumption of salt is dangerous even in the short term. So it is doubtful whether this particular result should be taken to significantly threaten my arguments concerning hunger.

kinds, and are therefore capable of forming the desire for food. Incentive learning will be needed even in humans to connect this desire to hunger, but once this is done, humans' motivation to obtain and consume food of any kind will be affected by their level of hunger, as long as they represent it as food *per se*.

The role of hunger in behavioural control is therefore indirect. Hunger acts as a stimulus, in the same way that perceived features of the environment do, in habitual control. In goal-directed control, the primary role of hunger is to influence the strengths of occurrent desires for foods, in cases in which incentive learning has taken place. This suggests that we should adopt a relatively narrow theory of desire, excluding basic drives, because it seems to be part of the typical philosophical conception that occurrent desires are intrinsically motivating.

5. Schroeder's Theory and Desires in Action

In the preceding three sections, I have presented my account of what desires are, and I have argued that hunger is not a desire. My arguments have been based on recent research in psychology and neuroscience. But as I have said, Timothy Schroeder's theory of desire draws on a similar body of empirical research, and differs from mine; it entails that basic drives are desires. Schroeder's theory may therefore identify a wider natural kind than mine, making it a particularly relevant alternative. It is also the best-known empirically-informed philosophical theory of desire. In this section I will explain how Schroeder's theory compares to mine, and present an objection to the theory. In section 6, I will turn to a further objection.

Schroeder's view is that what it is for an agent to desire some outcome *p* is for that agent to treat the occurrence of *p* as a reward.[7] His own statement of his theory is as follows:

> To have an intrinsic (positive) desire that *p* is to use the capacity to perceptually or cognitively represent that *p* to constitute *p* as a reward. (Schroeder 2004, p. 131)

> For an event to be a reward for an organism is for representations of that event to tend to contribute to the production of a reinforcement signal in the organism, in the sense made clear by computational theories of what is called 'reinforcement learning'. (Schroeder 2004, p. 66)

He also puts forward this theory in more recent work:

> To have an intrinsic appetitive desire that *p* is to constitute *p* as a reward. (Arpaly & Schroeder 2014, p. 128)

Schroeder's view, then, is that to desire an outcome is to have a certain kind of psychological disposition towards that outcome, which amounts to treating it as a reward

[7] Schroeder's talk of events as 'being rewards' is roughly equivalent, in the terminology that I have favoured, to saying that those events have positive reward values. However, one difference between us is that Schroeder gives a theory of reward, whereas I have not taken a position on this issue.

(and on his view, to that outcome's being a reward). He characterises what it is to treat an outcome as a reward by saying that this means being disposed to generate a more positive reinforcement signal when one represents that outcome than when one does not, other things being equal. And his view is that what it is for something to be a desire is for it to be the categorical basis of such a disposition. Since Schroeder takes phasic dopamine signals to be the only positive reinforcement signals in humans (and rats), for him human desires are the categorical bases of dispositions to produce positive phasic dopamine signals as a consequence of representing outcomes.

My theory is similar to Schroeder's in a number of respects. Because the states that I take to be desires contribute to measuring current levels of reward, they are likely to satisfy Schroeder's theory of desire. Schroeder and I agree that desire-acquisition works in a way that leads most humans to have desires for a wide range of different outcomes. And although he never explicitly mentions the distinction between occurrent and standing desires, Schroeder's theory does effectively accommodate this idea. For him, desires are dispositional states that persist in the long term, and are only active when their objects are occurrently represented. Thus on both my theory and Schroeder's, at any given time we are motivated only by a subset of our desires.

The main difference between my theory and Schroeder's is that his entails that standing basic drives are desires. This is because standing basic drives are by definition the bases of dispositions to produce reward signals. I have already presented two arguments against this view. First, creatures that lack the goal-directed system may possess basic drives; and second, basic drives do not have direct effects on action – instead, in the case of goal-directed action, their effects are mediated by incentive learning, and in the case of habitual action, their effects are no different from those of environment stimuli. The fact that *occurrent* basic drives do not have direct effects on action counts against the claim that *standing* basic drives are desires because real standing desires have occurrent forms that do have direct effects. If the suggestion in section 4 that there may be standing basic drives for which there are no corresponding occurrent drives is correct, this would also count against Schroeder's theory, because these states would be involved in action selection only by affecting the acquisition of desires and habits.

One possible response to these arguments is that while they show that basic drives are not states of the same kind that I am calling 'desires', they are members of another, broader natural kind, which better deserves to be called 'desire'. Even if Schroeder's theory identifies a natural kind, however, the narrower natural kind that I have identified better fits the common philosophical conception of desire, precisely because this conception takes desires to be intrinsically motivating, which neither standing nor occurrent basic drives appear to be. Furthermore, in section 6 I will argue that there is no natural kind encompassing both desires and standing basic drives.[8]

[8] Schroeder's stated view is that we have intrinsic desires to maintain homeostasis, and instrumental desires for food when blood sugar is low, for warmth when cold, etc. (2004, pp. 151-2). He therefore accepts that standing basic drives are desires. One awkward feature of this view is that it seems necessary for us to be able to learn about the reward values of nourishing foods even when we are not hungry, suggesting that we have a drive for food rather than for homeostasis.

I will now turn to a further problem with Schroeder's theory, which concerns his account of how desires causally contribute to motivation and action. According to Schroeder, desires influence our actions because when we contemplate a desired outcome, this causes a dopamine signal to be released, which in turn tends to make the performance of the action under consideration more likely (2004, pp. 115-8). This is a significant part of his overall account, because it implies a tight connection between the feature of desires in virtue of which they *are* desires – that is, their role in generating dopamine signals – and their ability to contribute to motivation and action. It also implies that basic drives and the states that I have called desires contribute to action in the same way, contradicting my argument of section 4. Both drives and desires dispose us to produce dopamine signals when we represent the outcomes that are their objects. However, this account of action is not an attractive one.

For one thing, the evidence Schroeder offers in favour of this account is not compelling. He cites three forms of evidence which are directly relevant to his view: studies suggesting that dopaminergic projections to the motor prefrontal cortex are necessary for maintaining motor intentions over time; the point that dopamine boosts action-selection by its effects on D1 and D2 receptors on medium spiny neurons in the striatum; and the fact that loss of dopaminergic activity causes impaired motion in Parkinson's disease (Schroeder 2004, pp. 116-118). One problem with all of this evidence is that while it does show that dopamine is necessary for motivation and action, this is far from sufficient to show that phasic dopamine signals are the means by which desires affect how we behave. More specifically, a widespread view is that while phasic dopamine signals represent RPEs and are used for reinforcement, it is *tonic* dopamine levels that affect motivation and the ability to initiate and control movements (which is lost in Parkinson's) (Niv et al. 2007, Schultz 2007). Tonic dopamine levels are the levels of dopamine release that obtain between phasic bursts. It could be that the role of tonic dopamine explains the evidence that Schroeder cites, while the effects of desires on action are mediated by a separate, non-dopaminergic mechanism.

Also, as I mentioned in section 3, there is an ongoing debate about the function of phasic dopamine signals, and neither of the two most prominent positions in this debate sits happily with Schroeder's account. On one hand, the neuroscientist Kent Berridge has defended the view that phasic dopamine signals have the function of generating motivation to bring about desired outcomes, but he sees this as an incompatible alternative to the view that phasic dopamine signals are for reinforcement (Berridge 2007). On the other hand, the orthodox view is that phasic dopamine signals are reinforcement signals, but do not play the role in motivation that Berridge proposes (Wise 2004, Balleine et al. 2008, Glimcher 2011). More generally, although the function of dopamine has been intensively studied, the evidence gathered so far presents a complex and contested picture, which does not clearly warrant Schroeder's conclusion.

Furthermore, it is hard to see how the mechanism Schroeder describes could work. The problem is that dopamine release is not targeted at particular groups of cells, and carries little information about how it is caused, so there is nothing except timing to distinguish one dopamine signal of a given strength from another. This means that the only way in which desires could be connected, via dopamine signals, to the correct actions is if the goal-directed system worked by considering actions in turn. In order to

choose between a number of available actions, the goal-directed system would have to represent them one at a time, use dopamine signals generated by desires at that time to associate these actions with reward values, and (presumably) store these reward values for later comparison. But this would be at odds with the point that not only action selection, but many other cognitive processes, seem to be constituted by competition between simultaneous coalitions of cortical activity. Concerning motivation and action specifically, a detailed account is available of how the basal ganglia facilitate the resolution of this kind of competition between cortical coalitions, which are connected to the basal ganglia by parallel partially-segregated closed loops (the striatum is part of the basal ganglia; for this account see Redgrave et al. 1999, Redgrave 2007). We have excellent reasons to think that the way action-selection happens involves the representation of possible actions simultaneously, rather than in turn.

This means that Schroeder's account of how desires contribute to motivation and action is in doubt. I have focused on the issue partly because it is intrinsically important, but also because Schroeder's account suggests that, contrary to the evidence from incentive learning, basic drives and desires contribute to goal-directed action selection in the same way. So the objections I have presented support my claim that desires and basic drives play distinct roles.

6. Standing Basic Drives as Psychological Primitives

Finally, a further reason to question Schroeder's theory is that standing basic drives, unlike desires, are *psychological primitives*. In this section, I will first outline the concept of a psychological primitive, then explain how this point counts against Schroeder's theory of desire.

According to Richard Samuels (2002), innateness claims in cognitive science can be most productively understood by taking 'innate' to mean 'psychologically primitive'. Samuels initially defines psychologically primitive traits to be those the acquisition of which is not correctly explained by psychological theories (2002, p. 246). For example, my capacity to feel pain is likely to be a psychological primitive, because a correct account of my acquisition of this capacity would presumably describe in the language of molecular biology and neurobiology how the relevant neural structures grew in the foetus that became me, rather than describing a cognitive process. One virtue of this account of innateness is that it entails that no psychological traits acquired through learning are innate, because learning is a psychological process; but it also correctly identifies that some psychological states that are not products of learning are non-innate, such as occurrent perceptual states. Samuels argues persuasively that this account captures what is at stake in prominent disputes between nativists and anti-nativists about various psychological traits.

However, Samuels notes that this initial account faces an objection. Psychological features that result from illness or injury are not innate, but also may not be susceptible to psychological explanations. To avoid this objection, he adds a further condition,

which is that traits are innate only if they are acquired 'in the normal course of events' (p. 259).[9]

Whether or not this account of innateness succeeds more generally, it does provide a useful way of making more precise the idea that there must be some outcomes which we are innately disposed to find rewarding. If there are systems that enable us to judge how much reward is provided by token outcomes, and to learn the reward values of outcome-types, then the actions of such systems are the only psychological processes by which we could acquire dispositions to find outcomes rewarding. But such systems could not exist without prior dispositions to find certain outcomes rewarding, because the reward values of outcomes are not properties of a kind that we can detect directly. This means that there must be some such dispositions which are not acquired through psychological processes, which is to say that they are psychological primitives. So standing basic drives are psychological primitives. Desires, in contrast, are not psychological primitives, provided anything remotely like the account of their acquisition given in section 3 is correct.

The problem for Schroeder's theory is that this contrast serves to make it doubtful that desires and standing basic drives are members of the same natural kind at all. Psychological primitives are a class of psychological structures (together with the dispositions and abilities they ground) which form the framework within which psychological processes take place. The human mind emerges from the way in which this set of structures interacts with its environment. Many psychological primitives, almost certainly including standing basic drives, are adaptations and are very close to universal in the population. These points entail that they are likely to take very different roles in scientific explanations from states which are contingent products of psychological processes, such as desires (although certain desires are also likely to be close to universal). It is therefore implausible that two psychological states, of which only one is a psychological primitive, could be members of the same natural kind unless they performed extremely similar functions. But we have already seen evidence that desires and standing basic drives play quite different roles in action selection, so we should be very sceptical of the claim that they are members of the same natural kind. We should consequently be similarly sceptical of Schroeder's theory.

7. Conclusion

In this paper, I have presented my own account of desire; argued against Timothy Schroeder's theory, which has similar aims and methods; and argued that hunger is not a desire. My own view is that desires are psychological states of the kind that play a particular role in goal-directed control: they keep track of the reward values of outcomes, and are used together with states tracking relationships between actions and outcomes to

[9] Samuels' account also faces a further potential objection, which is that the notion of innateness is used both in cognitive science and elsewhere (e.g. we can usefully distinguish innate and acquired components of the immune system; Mameli & Bateson 2011), yet his account is only relevant to cognitive science. This point does not diminish the usefulness of the notion of a psychological primitive for understanding the relationship between basic drives and desires.

determine how we act. These states are also used in assessing the reward values of current states of affairs, for the purpose of generating reinforcement signals.

In contrast, Schroeder's view is that what it is for a psychological state to be a desire is for it to play this latter role, in generating reinforcement signals. This view is questionable because it implies that standing basic drives are desires. I have argued in sections 4, 5 and 6 that basic drives do not affect motivation and action in the same way as desires, and that standing basic drives, unlike desires, are psychological primitives.

Hunger is not a standing basic drive, but it is also a mistake to think that hunger is a desire. This is partly because, as I show in section 4, behavioural experiments on rats show that hunger influences goal-directed behaviour only indirectly, through incentive learning. But in conclusion, I will briefly return to the points mentioned by Davis in the passage with which I began. Davis claimed that we can desire food without being hungry; that being hungry can explain the desire for food, but is not the only possible explanation; and that infants can be hungry while lacking the concept of food, but they cannot desire food. We can now see the force of these points more clearly. As some of the experiments mentioned in section 4 illustrate, it is possible to *acquire* desires for foods without being hungry, never mind merely experiencing occurrent desires. Occurrent desires can be generated by a number of factors, so while one way in which the desire for food might become occurrent is through hunger (in agents who have undergone appropriate incentive learning), there are other ways. An effective advertisement for a supermarket might make delicious food highly salient, and thus cause occurrent desires for food even in those who are not at all hungry. And finally, desires are formed when we represent outcomes at the same time as receiving positive reward signals. So an infant who lacked the concept of food would indeed be unable to acquire the desire for food, as opposed to desires for particular foods.

References

Adams, C. D. (1981). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology* 34B: 77-98.

Adams, C. D. & A. Dickinson (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology* 33B: 109-122.

Arpaly, N. & T. Schroeder (2014). *In Praise of Desire*.

Balleine, B. W. (1992). Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behaviour Processes* 18: 236-250.

Balleine, B., N. Daw & J. P. O'Doherty (2008). Multiple forms of value learning and the function of dopamine. In Glimcher (ed.), *Neuroeconomics*.

Balleine, B. W. & A. Dickinson (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37: 407-419.

Balleine, B. W. & J. P. O'Doherty (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology Reviews* 35: 48-69.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* 191: 391-431.

Bratman, M. (1987). *Intention, Plans and Practical Reason*.

Bray, S., S. Shimojo & J. P. O'Doherty (2010). Human medial orbitofrontal cortex is recruited during experience of real and imagined rewards. *Journal of Neurophysiology* 103: 2506-2512.

Brembs, B., F. D. Lorenzetti, F. D. Reyes, D. A. Baxter & J. H. Byrne (2002). Operant reward learning in *Aplysia*: Neuronal correlates and mechanisms. *Science* 296(5573): 1706-1709.

Butlin, P. (2016). *The Direction of Fit of Desire*. Ph.D Thesis. King's College, London.

Butlin, P. & D. Papineau (2016). Normal and addictive desires. In Segal & Heather (eds.), *Addiction & Choice*.

Carroll, L. (1895). What the tortoise said to Achilles. *Mind* 4: 278-280.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B* 362: 1585-1599.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences* 36(3): 181-204.

Corbit, L. H., J. L. Muir & B. W. Balleine (2001). The role of the nucleus accumbens in instrumental conditioning: evidence for a functional dissociation between accumbens core and shell. *The Journal of Neuroscience* 21: 3251-3260.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*.

Davis, W. A. (1986). The two senses of desire. In Marks (ed.), *The Ways of Desire*.

Dickinson, A., B. W. Balleine, A Watt, F Gonzales & R. A. Boakes (1995). Overtraining and the motivational control of instrumental action. *Animal Learning and Behaviour* 22: 197-206.

Dickinson, A. & G. Dawson (1988). Motivational control of instrumental performance: the role of prior experience of the reinforcer. *Quarterly Journal of Experimental Psychology* 40B: 113-34.

Foddy, B. & J. Savulescu (2010). A liberal account of addiction. *Philosophy, Psychiatry & Psychology* 17(1): 1

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences* 108(3): 15647-15654.

Griffiths, P. (2002). What is innateness? *The Monist* 85(1): 70-85.

Haber, S. N. & B. Knutson (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* 35(1): 4-26.

Hall, R. J. (2008). If it itches, scratch!. *Australasian Journal of Philosophy* 86(4): 525-535.

Hare, T. A., J. Malmaud, A. Rangel (2011). Focusing attention on the health aspects of food changes value signals in the vmPFC and improves dietary choice. *Journal of Neuroscience* 31: 11077-11087.

Heyes, C. & A. Dickinson (1990). The intentionality of animal action. *Mind & Language* 5(1): 87-103.

Holton, R. (2009). *Willing, Wanting, Waiting.*

Holton, R. & K. Berridge (2014). Addiction between compulsion and choice. In Levy, ed. *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience.*

Hull, C. L. (1943). *Principles of Behaviour.*

Hursthouse, R. (1991). Arational actions. *The Journal of Philosophy* 88(2): 57-68.

Jing, J., F. S. Vilim, C. C. Horn et al. (2007). From hunger to satiety: Reconfiguration of a feeding network by Aplysia neuropeptide Y. *Journal of Neuroscience* 27(13): 3490-3502.

Kennerley, S. & M. E. Walton (2011). Decision-making and reward in frontal cortex: Complementary evidence from neurophysiological and neuropsychological studies. *Behavioural Neuroscience* 125(3): 297-317.

Kirk, U., M. Skov, O. Hulme, M. S. Christensen & S. Zeki (2009). Modulation of aesthetic value by semantic context: an fMRI study. *NeuroImage* 44: 1125-1132.

Mameli, M. & P. Bateson (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B* 336: 436-443.

Neal, D., W. Wood, M. Wu & D. Kurlander (2011). The pull of the past: when do habits persist despite conflicts with motives? *Personality and Social Psychology Bulletin* 37: 1428-1437.

Niv, Y., N. D. Daw, D. Joel & P. Dayan (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology* 191(3): 507-520.

Niv, Y., P. Dayan & D. Joel (2006). The effects of motivation on extensively trained behaviour. *Leibniz Technical Report*, Hebrew University 2006-6.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Neuroscience* 34: 333-359.

Plassmann, H., J. O'Doherty, & A. Rangel (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience* 27(37): 9984–9988.

O'Doherty, J. P., J. Winston, H. Critchley, D. Perrett, D. M. Burt & R. J. Dolan (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia* 41(2): 147-155.

Rangel, A. & T. A. Hare (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology* 20: 1-9.

Redgrave, P. (2007). Basal ganglia. *Scholarpedia* 2(6): 1825.

Redgrave, P., T. J. Prescott & K. Gurney (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89: 1009-1023.

Rolls, E. T. & F. Grabenhorst (2008). The orbitofrontal cortex and beyond: from affect to decision-making. *Progress in Neurobiology* 86: 216-244.

Samuels, R. (2002). Nativism in cognitive science. *Mind and Language* 17(3): 233-265.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology* 80(1): 1-27.

Schultz, W. (2007). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience* 30: 259-288.

Sharot, T., T. Shiner, A. C. Brown, J. Fan & R. J. Dolan (2009). Dopamine enhances expectation of pleasure in humans. *Current Biology* 19(24): 2077-2080.

Schroeder, T. (2004). *Three Faces of Desire*.

Sescousse, G., J. Redouté & J-C. Dreher (2010). The architecture of reward value coding in the human orbitofrontal cortex. *The Journal of Neuroscience* 30(39): 13095-13104.

Skinner, B. F. (1938). *The Behavior of Organisms*.

Smith, M. (1987). The Humean theory of motivation. *Mind* 96: 36-61.

Smith, M. (1994). *The Moral Problem*.

Sulzer, J., R. Sitaram, M. L. Blefari, et al. (2013). Neurofeedback-mediated self-regulation of the dopaminergic midbrain. *NeuroImage* 175C: 176-184.

Sutton, R. & A. Barto (1998). *Reinforcement Learning: An Introduction*.

Tindell, A. J., K. S. Smith, K. Berridge & J. W. Aldridge (2009). Dynamic computation of incentive salience: 'wanting' what was never 'liked'. *Journal of Neuroscience* 29(39): 12220-12228.

Tricomi, E., B. W. Balleine & J. P. O'Doherty (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience* 29: 2225-2232.

Wise, R. (2004). Dopamine, learning and motivation. *Nature Reviews Neuroscience* 5: 483-494.

Yin, H. H., B. J. Knowlton & B. W. Balleine (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience* 19: 181-189.

Yin, H. H., B. J. Knowlton & B. W. Balleine (2006). Reversible inactivation of dorsolateral striatum enhances sensitivity to changes in action-outcome contingency in instrumental conditioning. *Behavioural Brain Research* 66(2): 189-196.

Yin, H. H., S. B. Ostlund, B. J. Knowlton & B. W. Balleine (2005). The role of dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience* 22: 513-523.